I. PRESERVACIÓN DIGITAL

Informe de situación

Preservación digital en 2009

Por **Miquel Térmens**

Térmens, Miquel. "Preservación digital en 2009".
Anuario ThinkEPI, 2010, v. 4, pp. 224-230

Resumen: *Se realiza un balance del estado de la investigación y el desarrollo de soluciones en preservación digital a nivel internacional en 2009, así como de las expectativas para 2010. Se presta una especial atención al seguimiento de las principales líneas de investigación y a los desarrollos de software específico, y se caracterizan los centros de investigación preeminentes a nivel mundial.*

Palabras clave: *Preservación digital, Biblioteca digital, Archivos digitales.*

Title: *Digital preservation in 2009*

Abstract: *The international state of research and development of digital preservation solutions in 2009 is assessed, as well as the expectations for 2010. We pay particular attention to the monitoring of the main research and development of specific software, and feature prominent research centres worldwide.*

Keywords: *Digital preservation, Digital library, Digital archives.*

LA PRESERVACIÓN de objetos digitales (una forma genérica de denominar cualquier documento o dato en formato digital) aún no cuenta en general con un corpus teórico asentado, metodologías aceptadas, normas de universal seguimiento y tecnologías implantadas.

Todos estos aspectos se encuentran en grados distintos de desarrollo, de tal forma que hoy en día, por decirlo de una forma simple, todavía no es posible adquirir un sistema de preservación digital llave en mano para una biblioteca, archivo, empresa o administración pública, excepto en el caso de grandes corporaciones.

Esta realidad no significa que no se esté trabajando o que aún no sea posible empezar a aplicar algunas metodologías, técnicas y soluciones para preservar documentos o datos de una determinada entidad. A continuación realizaremos un rápido repaso del estado de la disciplina en 2009 y de las novedades que se esperan para 2010.

1. Tendencias

Algunas de las líneas de investigación que parecen ya estar asentadas son el desarrollo de emuladores, la determinación de las propiedades significativas de los distintos tipos de documentos, la automatización de las migraciones de formatos y las metodologías de auditoría y certificación. A ellas se han sumado recientemente algunas nuevas líneas de trabajo que vamos a reseñar a continuación.

Existe una percepción generalizada de que los actuales sistemas de archivo de la Web se han de mejorar. Desde finales de la década de 1990, *Internet Archive*, no sólo se ha dedicado a archivar la Web pública mundial, sino que también ha desarrollado y ayudado a consolidar herramientas para el almacenamiento, indexación y recuperación de las páginas web, como por ejemplo *Heritrix*.
http://www.archive.org/

Estas herramientas y los procedimientos inherentes han sido de uso generalizado en todos los archivos web asociados en el *International internet preservation consortium* (*Iipc*) hasta el punto de que ha dado la impresión equívoca de que el archivo de la Web era un asunto solucionado. Esta visión dista mucho de la realidad: existen innumerables aspectos técnicos y organizativos mejorables o por resolver, como la sincronización de las capturas, las capturas incompletas de páginas

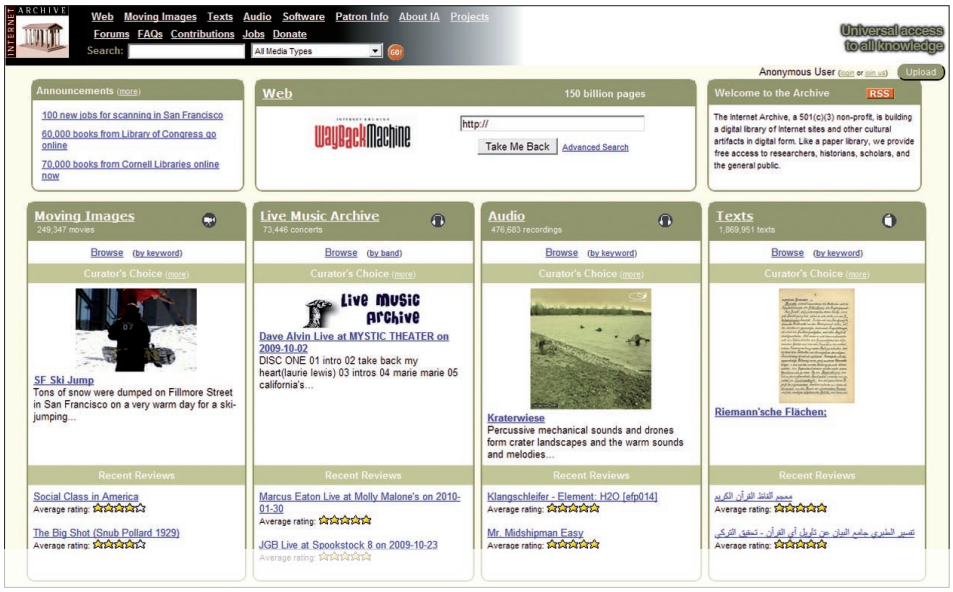


Figura 1. The wayback machine, http://www.archive.org

compuestas y la captura de webs dinámicos. Por ello, un análisis más profundo de la problemática relacionada con la preservación de la Web será el origen de una nueva generación de herramientas y técnicas en este sector.

"La preservación digital es todavía un área de trabajo poco conocida a nivel mundial"

Uno de los ámbitos en más rápida expansión es la preservación de las bases de datos científicas. En determinadas áreas del saber la investigación de base se está desarrollando cada vez más en red, en forma de consorcios y sobre la base de la colaboración internacional. La magnitud de este problema es enorme y unos pocos ejemplos nos lo mostrarán.

El detector *Atlas* del colisionador de partículas *LHC* del *Cern*, que entró en servicio en 2009, producirá 320 MB de datos por segundo cuando esté en pleno funcionamiento. El *Large synoptic survey telescope* (*Lsst*), que en 2014 entrará en servicio en Cerro Pachón (Chile), va a generar por cada noche de observación 15 terabytes (15×10^{12}) de imágenes celestes que pasarán a engrosar una base de datos de 60 petabytes (60×10^{15}).

Otros proyectos que en este momento ya están originando ingentes cantidades de información son el *Worldwide protein data bank* (*wwPDB*) y el *European molecular biology laboratory* (*Embl*). Y todos estos proyectos científicos internacionales generarán los datos cada vez a un ritmo mayor; su volumen y tasa de crecimiento superan los parámetros de las aplicaciones tradicionales y exigen la utilización de hardware y software específicos. Su preservación a largo plazo, o incluso su conservación a 5 ó 10 años vista, es una tarea que preocupa a los responsables de estas investigaciones.

Además de las actuaciones particulares iniciadas en cada uno de estos proyectos a favor de la preservación de sus resultados, empiezan a investigarse soluciones más globales y extensibles. Este es el objetivo a nivel europeo del proyecto *Parse.Insight*, financiado por la Unión Europea. Los EUA luchan en esta línea de actuación aún con mayor afán, ya que consideran que es una condición ineludible para mantener su primacía mundial en investigación. Una de las pruebas de este interés la tenemos en los 100 millones de dólares que durante 5 años a partir de 2010 va a invertir la *National Science Foundation* (*NSF*) en el llamado *Sustainable digital data preservation and access network partners* (*DataNet*), a llevar a cabo mediante un conjunto de proyectos novedosos. Uno ya adjudicado se encuentra *Data conservancy*, liderado por la *Johns Hopkins University*, con un presupuesto de 20 millones de dólares.

Dar con la solución para la preservación de los *datasets* científicos implica entre otras cosas poder almacenar grandes volúmenes de datos (del orden de petabytes) e interconectar las aplicaciones de alta capacidad de distintos centros de investigación. En este camino han aparecido dos términos: *data centres* y *cloud computing*.

Es necesario el uso de centros especializados en el almacenamiento masivo de datos (*data centres)*, aunque ello signifique almacenar los datos fuera de las universidades y empresas que los crearon. También se está avanzando rápido en el establecimiento de aplicaciones distribuidas, que aprovechen los recursos informáticos de distintos socios y que puedan gestionar y replicar datos en distintas localizaciones (*cloud computing*). Ya empieza a estar claro que el *cloud computing* tendrá en la preservación digital un sector importante de aplicación y negocio.

Varias razones aconsejan el uso de los recursos de *cloud computing*: la necesidad de disponer de mayores medios de procesamiento y almacena-

miento de información (no necesariamente en propiedad); la replicación de datos en almacenamientos remotos (por motivos de seguridad, recuerden el 11-S); e independizar las funciones de preservación de las propias de la gestión diaria. Una prueba de esta corriente la tenemos en el uso del sistema *iRODS*, creado por la *Universidad de Carolina del Norte*, por los archivos nacionales de los EUA (*Nara*) de manera que parte de los nuevos archivos federales electrónicos ya no están en la sede de la *Nara* en Virginia, sino "en la nube", entre distintos centros de datos de la red norteamericana.

Figura 2. Worldwide protein data bank, http://www.wwpdb.org

Una nueva problemática está llegando también a otro tipo de instituciones: las dedicadas a preservar la memoria del pasado, básicamente archivos y bibliotecas nacionales. Con los donativos de legados personales de políticos y otros personajes públicos empiezan a ingresar disquetes, cintas y discos con originales informáticos de estos personajes; los originales de las novelas ya no son manuscritos, ni siquiera hojas mecanografiadas, sino disquetes antiguos con ficheros grabados con programas ya desaparecidos.

"La *Universidad Técnica de Viena* destaca en la ingeniería del software aplicada a la preservación"

Empieza a ser un verdadero problema la identificación, conservación y consulta de estos nuevos fondos: los *eManuscripts*. Como ejemplo, cabe reseñar que la *British Library*, una de las instituciones afectadas, ha creado un departamento técnico específico para dar con soluciones para estas situaciones.

El control de los formatos técnicos de los ficheros a preservar es una estrategia clave en cualquier sistema de preservación y fundamentalmente en aquellos especializados en preservar documentos. Hasta el momento la iniciativa más importante bajo que de forma inicialmente descoordinada se estaba realizando desde los *National Archives* del Reino Unido, la *Library of Congress* y la *Harvard University*. En abril de 2009 se anunció el acuerdo de integración del *Global digital format registry* (*Gdfr*), de *Harvard*, con *Pronom*, de los *National Archives*, para crear el nuevo *Unified digital formats registry* (*Udfr*). El nuevo registro mundial de formatos cuenta desde sus inicios con el soporte, entre otros, de los archivos y bibliotecas nacionales de Canadá, EUA, Países Bajos y Reino Unido. Cuando entre en funcionamiento, a finales de 2010 o quizás en 2011, se convertirá en la primera herramienta técnica universal de soporte a las aplicaciones de preservación y en un ejemplo de cómo la cooperación internacional puede dar con soluciones aplicables a los distintos entornos de preservación.

El espectro de aplicación de la preservación digital cada vez se está abriendo más. La información textual y gráfica ya no es la única preocupación, sino que ahora también lo es la información sonora y la imagen en movimiento. Y en el año que la industria del cine se ha revolucionado con la llegada del 3D, no está de más recordar

que las nuevas películas ya rodadas en formato digital no tienen la vida asegurada.

Según los estudios de la propia academia de los Oscar (*The Academy of Motion Picture Arts & Sciences*), los ficheros originales de una película digital "normal" ocupan entre 2 y 10 petabytes y su coste de conservación es mucho mayor que el de una película tradicional: así, si el coste de conservación anual de un master tradicional se ha calculado en 1.059 dólares, en una película digital el coste sube a un mínimo de 12.514 dólares, sin contar y sin que se hayan resuelto los problemas técnicos que aparecerán a medio plazo (cambios de formatos...).

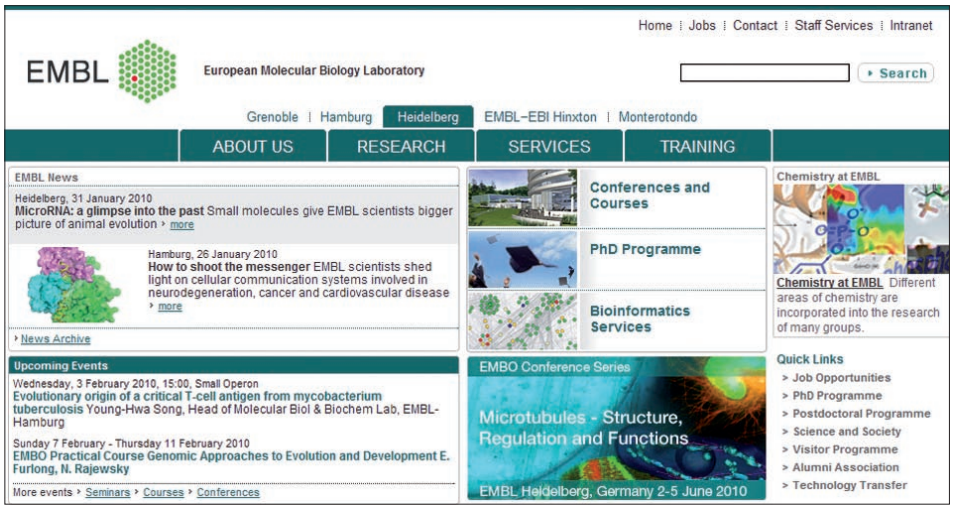


Figura 3. European molecular biology laboratory, http://www.embl.de

"El coste de conservación de una película digital es 12 veces mayor que el de un película en celuloide"

Nuevas oportunidades laborales

Un signo de la creciente aplicación de la preservación digital lo tenemos en el mercado laboral. En los países con más avances en este sector empiezan a publicarse ofertas de trabajo pidiendo perfiles del tipo *digital curator* o *web curator* y *digital preservation project manager*; los dos primeros son el conservador especializado en documentos digitales, que domina los formatos y las migraciones, mientras que el segundo es el director de un proyecto concreto, coordinando a un equipo multidisciplinar. En todos los casos se valora estar en posesión de conocimientos avanzados de indexación, esquemas de metadatos y organización de ontologías, así como de estructuración de información con XML.

También se empiezan a ofertar plazas para informáticos especializados, como *digital heritage archive developer* y *digital information systems analyst*, en las que será imprescindible el dominio de XML / XSLT, de lenguajes como Java y Perl, y de estándares como METS, MODS y OAI-PMH.

2. Software

2009 ha comportado interesantes novedades de software. El año se inició con el anuncio el 8 de enero de la compañía *ExLibris* del inicio de la comercialización de su producto *Rosetta*, específico para la preservación de objetos digitales, desarrollado a partir de la experiencia de la Biblioteca Nacional de Nueva Zelanda. Hasta este momento las funciones de preservación habían estado encomendadas a su software *Digitool*, un exitoso producto orientado en realidad a la gestión y acceso a colecciones digitales. Con la aparición de *Rosetta* se reconoce que la preservación digital es una actividad distinta a la gestión de un repositorio y portal de colecciones digitales y que requiere una tecnología propia.

De esta manera *ExLibris* se ha convertido en el tercer fabricante vendedor de software de preservación, después de *IBM* y de *Tessella*. Las tres empresas y sus respectivos productos se dirigen a mercados distintos: *IBM* a las grandes corporaciones privadas y a las administraciones públicas, *Tessella* a los archivos nacionales, y *ExLibris* a las bibliotecas nacionales y universitarias.

En mayo se hizo pública la fusión de *DSpace* y de *Fedora commons*, dos de los softwares de repositorios con más instalaciones, en el nuevo grupo *DuraSpace*. La fusión tiene como objetivo sumar los esfuerzos de las dos comunidades de desarrolladores con el fin de dar respuesta a las crecientes demandas de sus usuarios; entre éstas cabe destacar la necesidad de incorporar políticas de preservación en estos softwares, un punto

Figura 4. Ecdl 2009, http://www.ecdl2009.eu

limitaciones. *DuraSpace* también afrontará el reto de relacionar los repositorios con el *cloud computing*.

A nivel de herramientas de software este año también ha aportado novedades. *Droid*, que elaborada por los *UK National Archives* sirve para validar formatos, vio aparecer su v. 4 en julio de 2009: ahora ya puede trabajar con grandes discos a nivel de servidor. La alternativa a *Droid*, *JHove*, hasta ahora soportada por la *Harvard University*, reforzó su equipo con la incorporación de *California Digital Library, Portico* y *Stanford University*, con el fin de desarrollar la versión 2 en el periodo 2008-2010.

Por último se ha indicar que en los últimos meses se han puesto a libre disposición pública dos paquetes integrados de herramientas. El primero de ellos es *Roda 1.0*, el repositorio de preservación de archivos desarrollado por **Miguel Ferreira**, de la *Universidad do Minho* (Guimarães), por encargo de la *Dirección General de los Archivos* portugueses. El segundo es *Plato 2.1*, que forma parte de la iniciativa europea *Planets*; en este caso es una herramienta de planificación de proyectos de preservación preparada en la *Universidad Técnica de Viena*.

3. Núcleos de investigación y desarrollo

La investigación y el desarrollo de soluciones en preservación digital no es una actividad muy extendida, sino que por lo contrario se encuentra concentrada en unos pocos centros de investigación y países.

En Europa, el liderazgo de los Países Bajos es innegable, con su archivo y su biblioteca nacional al frente. Reino Unido destaca por las aportaciones de la *University of Glasgow*. Aunque con menores resultados, también se está actuando desde otros países: Alemania, Austria, Italia, Portugal, República Checa...

"*ExLibris* se ha convertido en el tercer fabricante vendedor de software de preservación, después de *IBM* y de *Tessella*"

Últimamente, tres nuevos focos están ganando peso y se están posicionando entre los líderes de la disciplina: el conjunto de la biblioteca y los archivos nacionales británicos, las universidades inglesas por impulso del *Jisc* (*Joint Information Systems Committee*) y la *Universidad Técnica de Viena*. En primer lugar, cabe prever que las acciones cada vez más visibles de la *British Library* y los *UK National Archives*, por ejemplo en el área de los *eManuscripts*, posiblemente se convertirán en ejemplos a seguir por sus homólogos de otros países.

En segundo lugar, el *Jisc* ha puesto la preservación digital como uno de sus máximos objetivos, como su exitosa trayectoria ya demuestra; esto significa que pronto tendremos a muchas universidades británicas trabajando en esta línea y aportando soluciones de aplicación inmediata.

En tercer lugar, los investigadores de la *Universidad Técnica de Viena* están destacando en la ingeniería del software aplicada a la preservación.

En España se debe citar el papel de cabecera de la *Biblioteca Nacional de España*, una institución que en lo técnico se ha visto claramente relanzada desde la llegada de **Milagros del Corral** como directora general. A nivel externo este impulso se está notando en la creciente incorporación a proyectos internacionales de cooperación, entre los que destacan *Europeana* y *Long term preservation (LTP)*. A nivel más práctico es de destacar que en diciembre de 2009 entró en servicio el depósito seguro de preservación *iArxiu*, creado por el *Consorci de l'Administració Oberta de Catalunya*, destinado a almacenar y preservar a largo plazo los documentos electrónicos de las administraciones públicas de Catalunya que se quieran adherir al mismo. Ofrece por tanto un servicio de custodia externa, un modelo de preservación digital que muy pronto tendrá una gran aceptación entre las administraciones y todavía más entre las empresas privadas.

Fuera de Europa sólo se detectan dos focos importantes de investigación: Oceanía y Estados Unidos. Australia y Nueva Zelanda son fuentes continuas de innovaciones, siempre de carácter aplicado, generadas por sus bibliotecas y archivos nacionales, por determinadas universidades y también por gobiernos regionales, como el de Victoria, en el marco de la administración electrónica.

Estados Unidos son líderes en preservación por tres razones:

1. Cuentan con numerosos núcleos de investigación muy sólidos;
2. Existe una capacidad de trabajo colaborativo entre estos núcleos que les permite sumar esfuerzos y realizar proyectos a gran escala; y
3. Algunas instituciones federales, como la *Library of Congress* y la *Nacional Science Foundation*, o de promoción de la investigación como la *Andrew W. Mellon Foundation*, están realizando un papel de incentivación y regulación de la investigación con gran acierto.

La preservación digital es todavía un área de trabajo poco conocida a nivel mundial, en parte debido a su corta existencia y al reducido número de especialistas que se dedican a ella. Ello provoca, entre otras consecuencias, la falta de unos canales formales y específicos de comunicación científica. Esto es especialmente cierto en el caso de los artículos, que tienden a repartirse entre las publicaciones de biblioteconomía, archivística y, sobre todo, de ingeniería informática. En los congresos, en cambio, se están consolidando unas pocas convocatorias. El congreso más importante es *iPRES*, que alterna su celebración anual entre Estados Unidos y Europa (la edición de 2010 tendrá lugar en septiembre en Viena). También destacan la *European conference on digital libraries* (*Ecdl*), en Europa, y la *Joint conference on digital libraries* (*Jcdl*), en Estados Unidos.

4. Bibliografía

Becker, Christoph; **Kulovits, Hannes**; **Rauber, Andreas**; **Hofman, Hans**. "Plato: a service oriented decision support system for preservation planning". En: *Proceedings of the 8th ACM/IEEE-CS Joint conf on digital libraries, JCDL'08*, June 16–20, 2008, Pittsburgh, Pennsylvania, USA. ACM, pp. 367-370.

Bock, Nicholas. "Preserving the data harvest". *Simmetry. Dimensions of particle physics*, 2009, v. 6, n. 6, pp. 18-22.

Dappert, Angela; **Farquhar, Adam**. "Significance is in the eye of the stakeholder". M. Agosti et al. (eds.): *Research and advanced technology for digital libraries, 13th European conf., ECDL 2009*, Corfu, Greece, Sept. 27 - Oct. 2, 2009. *Proceedings, Lecture notes in computer science*, v. 5714, pp. 297-308.

The digital dilemma. Strategic issues in archiving and accessing digital motion picture materials. Academy of Motion Picture Arts and Sciences, 2007, 74 pp.

Térmens, Miquel. "Investigación y desarrollo en preservación digital: un balance internacional". *El profesional de la información*, 2009, v. 18, n. 6, pp. 613-624.

Informe anual

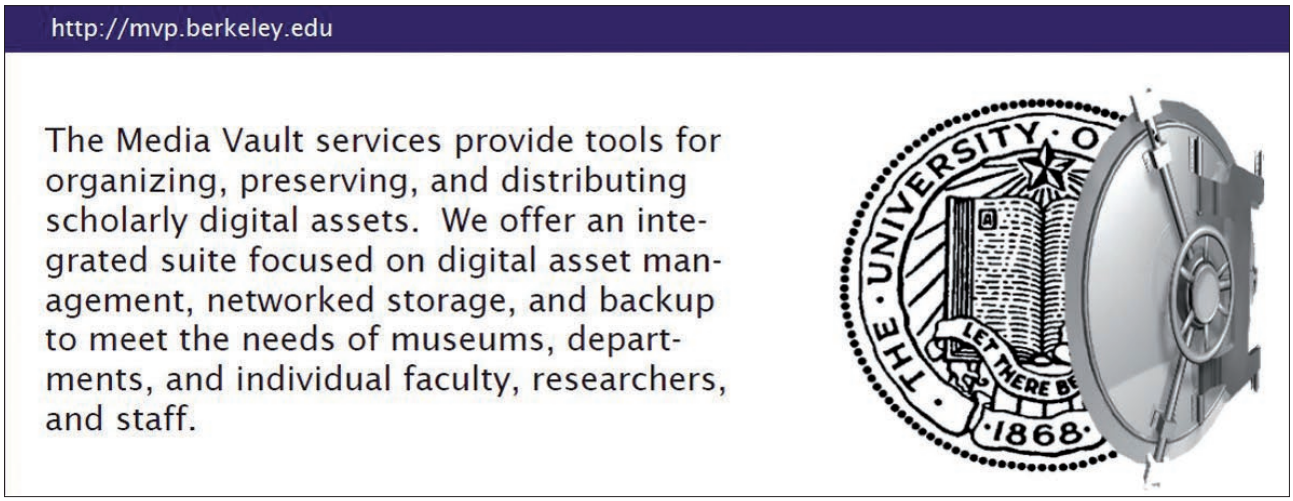
Media Vault Program de Berkeley

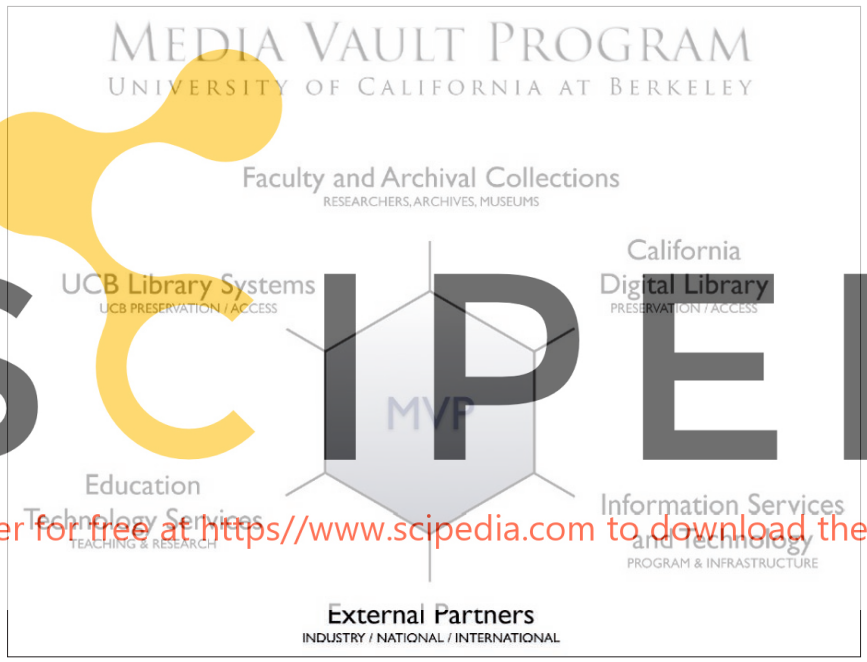
MVP interim report. Sept. 2009
http://mediavault.wordpress.com/tag/interim-report/

El *Programa Media Vault* (*MVP*) es una aproximación interdisciplinar para promover el archivo digital de los flujos de trabajo académico, conservar la información digital, ofrecer un auto-servicio de gestión de las colecciones digitales, y asegurar un entorno digital de apoyo a la investigación y a la enseñanza, y de servicio público.

El principal problema es la falta de recursos en todo el campus de Berkeley para abordar de forma adecuada nuestra dependencia cada vez mayor de los medios digitales.

Algunas conclusiones:





– El problema es grande y hay que buscar soluciones coherentes. Proveedores de servicios y técnicos tenemos que trabajar juntos y armonizar los esfuerzos en la mayor medida posible.

– El problema es manejable, se puede avanzar gradualmente. Existen medidas pragmáticas y relativamente baratas que se pueden aplicar de inmediato, que redundarían en importantes beneficios.

– Algunas necesidades son básicas: un lugar seguro para poner las cosas y una manera fácil de compartirlas. Tener un lugar seguro para guardar los datos de investigación daría mucha tranquilidad al personal.

– Otras necesidades son complejas: la preservación digital y el acceso permanente es difícil a largo plazo. La transferencia de responsabilidades del creador al *curator* (el encargado de guardar los datos) trae consigo una gran complejidad debido a los requisitos que normalmente se presentan para asegurar la transición. Hay que ser pacientes y generosos con la comunidad de usuarios y darse cuenta de que la complejidad de este ámbito es un obstáculo para la adopción del *MVP*.

– Hay pocos incentivos *para hacer lo correcto*. Es necesario concienciar, y comunicar las mejores prácticas.

– Existe deseo de aprender y compartir. Una de las fortalezas de trabajar en un ambiente académico es el deseo general de [illegible] y de tolerancia a la imperfección.

[...]

– Son posibles las soluciones comunes. Al centrarnos en el flujo de trabajo y en el ciclo de vida, los puntos conflictivos se revelan iguales para la mayoría de los usuarios. Hay departamentos con miles de imágenes, otros tienen menos archivos pero quizá necesitan compartirlos mucho más. La escala es relativa.

[...]

Todo debe funcionar en auto-servicio. Los usuarios tienen diferentes necesidades, capacidades de pagar o de contribuir. No hay una escala móvil entre los ricos que pueden pagar los servicios completos y los que no pueden. De hecho, el auto-servicio, es decir, auto-empoderamiento, debe ser un objetivo. En la medida de lo posible, la empresa de investigación debe ser a la vez independiente y totalmente compatible. El auto-servicio es clave para problemas de escalabilidad humanos para los proveedores, que se traduce en menores costos y mayor capacidad de respuesta.